Z - 149E
C. 2.

AN OVERVIEW OF SMALL AREA ESTIMATION TECHNIQUES

March, 1985

J. Dumais
S. Earwaker
J.-F. Gosselin
D. Paton
K.P. Srinath
R. Verma

CHSM 85-074E

TABLE OF CONTENTS

1. Introduction

   1.1 Purpose
   1.2 Data Sources
   1.3 Estimation Techniques

2. Survey or Design Based Estimation

   2.1 Introduction
   2.2 Description and Discussion

3. Methods Combining Survey and Administrative or Census Data

   3.1 Introduction
   3.2 Synthetic Estimation
   3.3 SPREE Estimation
   3.4 Sample Regression Method
   3.5 Composite Estimation

4. Methods Combining Census and Administrative Data

   4.1 Introduction
   4.2 Component Methods
   4.3 Non-Component Symptomatic Methods
   4.4 Regression Symptomatic Technique

SMALL AREA ESTIMATION TECHNIQUES

1. INTRODUCTION

1.1 PURPOSE

Methods of producing estimates for small areas or small domains have been of interest to demographers and statisticians for many years, and as a result a wide variety of techniques have been developed. With the growing demand for small area estimates within Statistics Canada, it is becoming more important that analysts be able to select techniques appropriate to their application from the broad array of those available.

This paper provides a non-technical description of a variety of small area estimation techniques but by no means represents an exhaustive review of all techniques available. It was prepared to assist persons who may wish to acquire a basic understanding of these methods, highlighting their data requirements, advantages and disadvantages. A bibliography appeared in Survey Methodology IX, 2, and would be useful to readers who may wish to explore the subject further.

1.2 DATA SOURCES

The three principal sources of data used to produce small area estimates are censuses, sample surveys and administrative records. For small area estimation, each of these sources has strengths and weaknesses, some that are shared with the others, and others that are not.

Censuses have traditionally been the primary source of small area data. However, because of their cost, censuses tend to be carried out infrequently and hence cannot be used to monitor local area conditions.

Sample surveys, on the other hand, are often conducted more frequently, and so are more useful for monitoring short-term changes in the population. They also typically collect a large number of data items, allowing a wider range of estimates to be made. However, sample survey based estimates for small areas often have unacceptably high sampling errors.

Administrative data, like census data, have good small area potential. They are often available at shorter intervals but typically contain a limited range of variables. They may be subject to limitations such as differences in coverage or concepts and other data quality problems.

1.3 ESTIMATION TECHNIQUES

Small area estimation techniques can generally be divided into three groups on the basis of the source or sources of data that they use. The first group consist of the strictly survey based estimators, the second, of those that combine survey with one or both of census and administrative data, and the third consist of those techniques that use both census and administrative data.

1.3.1 Survey or Design Based Estimators

This group includes all the traditional sample survey based estimators. These are approximately unbiased[1], and the major sources of error are measurement and sampling error.

1.3.2 Methods Combining Survey and Administrative or Census Data

These methods have been developed in an attempt to obtain more accurate estimates by using auxiliary information from census or administrative sources in concert with survey data.

The techniques make implicit or explicit use of models and are therefore subject to model error as well as sampling and measurement errors.

Methods in this group that will be discussed in this paper are:

Synthetic Estimation

---

[1] In practice, sample survey estimates rely to some extent on modelling, eg imputation, ratio-estimation.

Structure Preserving Estimation
Sample Regression
Composite Estimation

1.3.3 Methods Combining Census and Administrative Data

These methods have been developed extensively by demographers. Census data serve as benchmarks and are updated using auxiliary data available at the small area level. Typically, administrative data are used as the auxiliary information.

These techniques make explicit use of models, but no use is made of survey data. Thus the sources of error in the estimates are model error and measurement error.

Methods in this group that will be discussed are:

Component Methods
Non-Component Symptomatic Methods
Regression Symptomatic Technique

2. SURVEY OR DESIGN BASED ESTIMATION

2.1 INTRODUCTION

It is sometimes possible by applying standard estimation methods, to produce estimates for areas other than the ones for which the survey was designed, provided the sample size falling into the area is sufficiently large. The estimators described in this section are those which are appropriate to this situation.

2.2 DESCRIPTION AND DISCUSSION

Let a denote the small area and let:

Y_a: total for the variables of interest for area a

N_a: total number of units in area a

X_a: auxiliary information for area a.

The following estimators may be constructed:

2.2.1 Direct Estimator

This is obtained by applying the estimation procedure appropriate to the design to the sample observations falling into the small area a. The estimator takes the form

$$\hat{Y}_a^{(1)} = \sum_{i \in a} W_i y_i$$

where W_i and y_i denote the sample weight and the observation associated with the i-th sample unit, the sum being taken over all units in small area a.

This estimator is design-unbiased and the small area estimates will be consistent with survey estimates produced for higher levels of aggregation. The main disadvantage is the possible high level of sampling error associated with such estimates. Realistically the method can only be used in cases where the sample size in the small area is sufficiently large to provide estimates with acceptable level of sampling error. It should be recognized that these estimates may be contitionally biased.

2.2.2 Ratio Estimators

In the situation where auxiliary information such as X_a or N_a, is known from external sources and can be estimated from the sample, we can construct two possible ratio estimators:

$$\hat{Y}_a^{(2)} = \frac{\hat{Y}_a^{(1)}}{\hat{N}_a^{(1)}} \cdot N_a$$

$$\hat{Y}_a^{(3)} = \frac{\hat{Y}_a^{(1)}}{\hat{X}_a^{(1)}} \cdot X_a$$

These estimators will produce a reduction in the sampling variance compared to the Direct Estimator provided the variable Y is positively correlated with the auxiliary variables N or X. Ratio estimators however are biased and may become very unstable for very small samples. Extra caution must therefore be exercised in using these estimators.

It should be noted that ratio estimation may introduce inconsistencies within estimated totals at higher levels of aggregation.

2.2.3 Regression Estimators

Regression estimators like ratio estimators can be used to improve the direct estimator in the case where auxiliary information X which is correlated with the study variable Y is available. Examples are:

$$\hat{Y}_a^{(4)} = \hat{Y}_a^{(1)} + \hat{\beta} \, (X_a - \hat{X}_a^{(1)})$$

$$\hat{Y}_a^{(5)} = \hat{Y}_a^{(1)} + \hat{\beta}_a \, (X_a - \hat{X}_a^{(1)})$$

where $\hat{\beta}$ and $\hat{\beta}_a$ are regression coefficients estimated from the entire sample or the portion of the sample falling into the domain respectively. These are special cases of a very general regression estimator developed by Särndal (1981, 83, 84) which can be constructed for any number of auxiliary variables.

Regression estimators share many of the characteristics of Ratio estimators. They are essentially unbiased for large samples but may become unstable (particularly $\hat{Y}_a^{(5)}$) as the sample size decreases. They will produce a reduction in sampling error compared to the direct estimator if the variables Y and X are correlated. They may also introduce inconsistencies in the estimates.

3. METHODS COMBINING SURVEY AND ADMINISTRATIVE OR CENSUS DATA

3.1 INTRODUCTION

The methods described in this chapter use sample survey data and auxiliary data sources (e.g. administrative records or census data).

In a typical application of these techniques, current survey information is available, but the areas for which we desire estimates are too small to be accomodated by the survey program.

The following methods rely upon modelled relationships between the survey data and the auxiliary data. It is an attempt to reduce the high levels of sampling error associated with survey estimates for such small areas.

Four examples of this type of estimation will be discussed here. These methods make explicit use of statistical models and are subject to model error whenever the underlying assumptions are not met.

Since these methods make use of current survey data, they are not subject to out-of-dateness. This may provide a distinct advantage over static models such as symptomatic regression (see 4.3 below), when either the variable of interest or the model, or both, are unstable over time.

3.2 SYNTHETIC ESTIMATION

3.2.1 Introduction

Synthetic estimation is a procedure whereby small area estimates are derived by combining survey estimates available for larger domains with Census or other data sources such as administrative records which pertain to the small area for which estimates are required.

3.2.2 Description

Again let 'a' denote a small area for which an estimate is to be produced. Suppose the population is subdivided into mutually exhaustive groups denoted by 'g'. Let X_{ag} denote auxiliary information available for group 'g' and area 'a', and let $\hat{Y}_{.g}$ be a sample survey estimate of the variable of interest for group 'g' across all areas. Then the synthetic estimate may be expressed as

$$\hat{Y}^{*}_{a.} = \sum_{g} \hat{Y}_{.g} \frac{X_{ag}}{X_{.g}}$$

This quite simply involves the re-weighting of survey estimates applicable to larger domains with the use of auxiliary information at the small area level.

3.2.3 Discussion

The main advantage of the synthetic estimator is that it produces estimates with relatively low sampling error. This stems from the fact that the estimator is a function of design-based estimates for relatively large domains (i.e., $\hat{Y}_{.g}$'s).

The procedure however could suffer from serious bias. The underlying assumption is that within domains g, the auxiliary information (X_{ag}) can be used to allocate the design-based estimates to small areas i.e.

$$\frac{Y_{ag}}{Y_{.g}} = \frac{X_{ag}}{X_{.g}}$$

In the case where auxiliary variable corresponds to population counts for subgroups of the population (eg., age-sex groups) this is equivalent to assuming that the groups are homogeneous with respect to the variable of interest and that variations in the variable of interest can be fully

explained by variations in the composition of the population with respect to groups 'g'. These are relatively strong assumptions for most applications and the procedure will often tend to pull small area estimates towards means for the larger areas.

However the procedure can prove very useful in situations where the auxiliary information such as administrative data is highly correlated with the variable of interest.

From an operational point of view the procedure is very simple to apply.

3.3 SPREE ESTIMATION

3.3.1 Introduction

Structure PREserving Estimation (SPREE) is a generalization of Synthetic Estimation. It involves a fuller use of survey estimates together with auxiliary information.

3.3.2 Description

The methods makes use of a technique called Iterative Proportional Fitting (IPF) also referred to as Raking Ratio Estimation Procedure. The method has several variations but can be described in the context of a specific application.

It is assumed that survey estimates for characteristics of interest are available at some higher level of aggregation, but not reliable at the small area level 'a'. Let $\{\hat{Y}_{.g.}\}$ and $\{\hat{Y}_{..h}\}$ represent survey estimates corresponding to a breakdown into classes defined by two distinct variables. For example 'g' could represent age-sex groups while 'h' may correspond to marital status categories. It is assumed that auxiliary information $\{X_{agh}\}$ is available at the small area level from an alternative source (eg census or administrative records).

Using $\{\hat{Y}_{.q.}\}$ and $\{\hat{Y}_{..h}\}$ as marginal totals, the method then involves adjusting the matrix $\{X_{aqh}\}$ using the IPF procedure:

a) First, the ratio $\hat{Y}_{.q.}/X_{.q.}$ is calculated and every element in the matrix for category 'q' is multiplied by this ratio. This results in an adjusted matrix $\{\hat{Y}^{(1)}_{aqh}\}$ which is consistent with $\{\hat{Y}_{.q.}\}$.

b) The same procedure is applied with respect to the second set of classes with ratios $\{\hat{Y}_{..h}/\hat{Y}^{(1)}_{..h}\}$. This result in an adjusted matrix $\{\hat{Y}^{(2)}_{aqh}\}$ which is consistent with survey estimates $\{\hat{Y}_{..h}\}$.

c) Steps (a) and (b) are then repeated using the updated matrix until the process converges.

The SPREE estimates are then obtained using the final matrix.

It is worth noting that the first adjustment (a) of this procedure corresponds exactly to the synthetic estimator with respect to the corresponding set of classes.

3.3.4 Discussion

The sampling error associated with SPREE estimators is relatively low although it may in some cases be higher than the synthetic estimator. The primary advantage of the SPREE method is that it allows the use of more than one breakdown of survey estimates (two in the above descriptions but this can be generalized).

The assumption underlying this method is that except for the adjustments made using survey estimates, the matrix of auxiliary information closely reflects the structure of the variable under study at the small area level and hence can be used to "allocate" the survey estimates to small areas.

In some applications the matrix is obtained from a previous census which is then updated using current survey estimates. This involves the relatively strong assumption that the relative level of the estimate within classes 'q' and 'h' has not changed since the previous census at the small area level except for what can be explained using the survey estimates at higher levels of aggregation.

Other types of applications involve the use of administrative data as auxiliary information. This can be quite powerful if the information obtained from administrative records is highly correlated with the variable of interest and is current. In this particular case, the estimation procedure can be viewed as adjustments to compensate for previous differences/deficiencies in concepts and coverage of administrative records relative to the statistical concept underlying the survey estimates.

SPREE estimation is relatively more complex to apply than synthetic. However generalized programmes are available to carry out the procedure.

3.4 SAMPLE REGRESSION METHOD

3.4.1 Introduction

In the sample regression approach, small area estimates are calculated from a regression equation. The equation itself is constructed and fit on the basis of current sample survey estimates which exist for small areas (sampled) from the larger areas of interest.

3.4.2 Description

Estimates are calculated on the basis of a linear model of the form

$$Y = X \beta + \varepsilon$$

where Y is the dependent variable,
X is a set of independent variables,
β is model coefficient

and ε is a random error.

The method is applied as follows:

1. Values of 'Y' are estimated for units sampled from the survey frame (e.g. primary sampling units) using the survey design-based estimation procedure.

2. Symptomatic (auxiliary) information is collected to obtain values of the independent variables, 'X', for the same units from the survey frame, and for small areas of interest.

3. Least squares regression is used to estimate the 'β' coefficient for the equation, based upon the observations of 'X' and 'Y' available for sampled units.

4. Values of the symptomatic variables 'X' are used in the equation to produce estimates of 'Y' for small areas of interest.

The method was originally developed by Ericksen (1974) for the purpose of estimating population of small areas.

3.4.4 Discussion

Users should consider the nature of their particular estimation problem very carefully when applying this method. For example, some knowledge of the measurement of 'Y' values and its error may suggest some procedure other than least squares for fitting the regression equation. Also, all variables used in the regression may be written in ratio form to reduce the variability and skewness of their distributions. For example variables can be expressed as a ratio of change (or difference) from the previous Census values (as in Ratio-Correlation and Difference-Correlation methods). Estimates of change produced from such a model would be multiplied by (or added to) the previous Census values to obtain current small area estimates of level.

The advantages of the method are:

1. It makes use of current information on Y to build an up-to-date estimation model.

2. It is very flexible.

3. The model is relatively robust to variation in survey estimates.

The disadvantages of the method are:

1. It relies heavily upon the model. Local variations in the relationships between Y and the X variables may not be reflected in the estimation procedure. Thus variations in Y must be completely and uniformly measured by the X variables. This may be very reasonable for some situations such as those examined by Ericksen. For example, estimates of population change may be estimated from births, deaths and estimates of migration flows.

2. The method relies upon the assumption that relationships between Y and X's established for sampled units (e.g. primary samplings units) will also hold for non-sampled units.

3. The symptomatic information must be available for units on the survey frame. Sometimes such information from auxiliary sources is not readily available for survey units and must be converted or adjusted in some way to obtain correspondence.

3.5 COMPOSITE ESTIMATORS

3.5.1 Introduction

Composite estimation relies on a combination of two or more estimation methods.

3.5.2 Description

Most composite estimators which have been proposed use a linear combination of a design-unbiased estimator $\hat{Y}_1$ and a model-based estimator $\hat{Y}_2$ of the form

$$\alpha_1\hat{Y}_1 + \alpha_2\hat{Y}_2$$

For example $\hat{Y}_2$ may be a sample regression estimate or may in fact be a synthetic or SPREE estimate. The objective is to produce an estimator which has lower error than any of the single components. The problem then is to select the appropriate weights α_1 and α_2.

It is possible to derive a mathematical expression for deriving optimal values of α_1 and α_2. Generally such expression involves terms that are difficult or impossible to estimate. One approach suggested by Schaible et al (1977) weights each component in proportion to the inverse of its squared error. Variation of this principle lead to the so called Empirical Bayes estimators.

Another approach developed at Statistics Canada (Drew, Singh, and Choudhry 1982) refered to as the Sample Dependent Estimator uses weights based upon the sample yield in the small area of interest. The principle behind the method is that when the sample yield is sufficiently large that a traditional design-unbiased estimator will be reliable, then the full weight should be given to this component. When the yield is insufficient then a weighted combination of the design-unbiased and synthetic estimators is used with gradual reliance on the synthetic component as the sample size decreases.

3.5.3 Discussion

Composite approaches are still relatively new and are under active research. Proponents claim the following advantages to these approaches:

1) The methods yield lower mean squared error than the M.S.E.'s for components alone.

2) The methods allows the use of available information from both within the small area of interest and auxiliary information either from administrative sources or sample data from adjacent areas.

Some apparent disadvantages are:

1) The methods are more complicated and costly to develop than any single component estimator, since at least two must be developed.

2) The optimal weights for combining estimators are never known, but must be estimated and are subject to error. Certain methods may be impractical or impossible to implement in particular circumstances due to the difficulty of estimating the correct weights.

4. METHODS COMBINING CENSUS AND ADMINISTRATIVE DATA

4.1 INTRODUCTION

These methods have been primarily developed by demographers for estimation of population. The administative data that may be available include births, deaths, tax returns, school enrollment, family allowance recipients, telephone installations, hydro connections, occupancy permits issued, motor vehicle licenses and voter registration etc., and are often referred to as symptomatic data. They are all symptomatic of population size or population change, but each in its own way.

Births are directly related to the change in the population at the age 0, whereas deaths affect each age of the population. School enrollments as well as family allowance recipients are indicators of children under 17 years of age.

Telephone and hydro connections and disconnections, reflect migration.

In general, several symptomatic variables are used in conjunction with census data to produce post-censal estimates. Generally, the census is used to provide a benchmark.

The quality of the estimate of population change using the symptomatic data, is dependent on the coverage and quality of the administrative files for population estimation. A file which seems to be permanent and covers a large extent of population is always preferred for estimating population. In practice, no administrative file is perfect. However, births and deaths, family allowance and tax returns have been found to be useful in Canada and the U.S.A.

Component, non-component and regression symptomatic methods, are three methods described in the following sections that use the relationship between the census counts and symptomatic variables, and may be useful for small area estimation.

4.2 COMPONENT METHODS

4.2.1 Introduction

Component methods depend on dividing the population into components of change, estimating the size of the components separately, and combining them to estimate the total. For small area estimation, these techniques require that comprehensive and detailed data sets be currently available at the small area level.

4.2.2 Description and Discussion

a) Balancing Equation

Estimates are derived using the following balancing equation:

$$P_a(t+1) = P_a(t) + B_a(t,t+1) - D_a(t, t+1) + NM_a(t,t+1)$$

where

$P_a(t+1)$ = Estimated population of a'th small area at time "t+1"

$P_a(t)$ = Census count of a'th small area at time "t"

$B_a(t,t+1)$ = Births occurring in a'th small area between time "t" and "t+1"

$D_a(t,t+1)$ = Deaths occurring in a'th small area between time "t" and "t+1"

$NM_a(t,t+1)$ = Net migration for a'th small area between time "t" and "t+1"

The quality of the resulting estimate is a function of the quality of the component estimates. Vital records are virtually complete and the only significant source of error in the vital records is geographic mis-classification, which may introduce bias in the estimates at the small area level. Migration is the component most difficult to estimate; generally it is estimated using symptomatic variables.

b) Cohort-Survival

The cohort-survival model is used both in estimating and projecting population. Population is disaggregated into male and female age cohorts. Each age cohort spans five years or one year age interval. Age-specific death rates are obtained for each cohort. Age specific birth rates are applied to female cohorts to estimate the total number of births. Each cohort group is then aged forward towards the estimation year, with age specific mortality. Finally, population is obtained for the estimation year by adding or subtracting the migration component to the specific cohort group. The advantage of this technique is the excellent detail it provides in projecting future demand for age-specific needs, such as schools, jobs or service for the elderly. It is a fairly accurate forecasting technique when migration is either known or negligible.

c) Composite Methods

Composite methods divide the population into subpopulations, and estimate each size separately. The subpopulations are often age classes. Estimates are obtained using the symptomatic data most closely related to each of the subpopulations. For example, births are used to estimate the population under 5 years old; school enrollements for ages 5-17; health care files or pension files for ages 65 and over and various administrative files to measure the 18-65 age group (Boque and Duncan, 1959).

This method has been used in both Canada and the United States and has proven to be a valuable technique for small area population estimation, but not without disadvantages. In the case of the application outlined in the last paragraph it is known that migration will be underestimated when it is based on school enrollments and errors in estimating the female population 15-44 will be compounded when the estimate of the male population is made, since it is based on the female population.

4.3 NON-COMPONENT SYMPTOMATIC METHODS

4.3.1 Introduction

These methods attempt to model the net change in population by making use of auxiliary or symptomatic variables that are correlated with population size such as births and deaths, tax returns, family allowance recipients, hydro connections.

4.3.2 Description and Discussion

a) Vital Rates

The Vital Rates Method provides estimates of population for small areas using the following sets of input data:

- Crude birth rate (or crude death rate) for the census year for each small area

- Number of births (or deaths) for the estimating year for each small area

- Provincial crude birth rate (or crude death rate) for estimating year as well as for census year.

In order to estimate the crude birth rate (or crude death rate) of the small area for the estimating year, the ratio of the provincial crude birth rate (or crude death rate) for the estimating year to that for the census year is applied to the crude birth rate (or crude death rate) of the small area in the census year. The population for the small area is then estimated as the number of births for the area are divided by its estimated crude birth rate. In the same way, a second population estimate can be derived using deaths and the estimated crude death rate.

This method is simple but its accuracy depends on the relationship between the small area's birth rate and the entire population's birth rate remaining constant between the census and estimating years.

b) Trend Extrapolation

This method uses the historical growth pattern to estimate the future growth pattern. It deals with the net effects of births, deaths and migration rather than the individual components. After plotting past population growth, one fits a curve and extends it into the future. Linear or non-linear regression formulae may be used, depending on whether past growth shows linear, exponential or logistic patterns.

For estimation purposes, population can be related to historic trends in employment, school enrollments, housing units, motor vehicle registrations or other symptomatic data.

This method is simple and reliable for slow-growth rates and short-term estimations. The primary disadvantage of this method is the lack of component data. Relationships among births, deaths and migration are non-available, and these components may have different trends of their own that will lead to a different net growth pattern in the future.

c) Ratio Trend

This method assumes that the small area makes up the same proportion of the larger area as it did in the past and is useful when accurate current estimates are available for the large area and accurate counts are available for the small and large areas for some previous time.

This method is simple, but risky in the assumption that historic relationships will hold in the future.

d) Proportional Allocation

This method relates the proportions of a small area to be a larger area with respect to population and to a symptomatic variable. It is assumed that the changes in these proportions from year to year are the same for both the population and for the symptomatic variable. The following equation is used:

$$\frac{\frac{P_a(t+1)}{P_\cdot(t+1)}}{\frac{P_a(t)}{P_\cdot(t)}} = \frac{\frac{S_a(t+1)}{S_\cdot(t+1)}}{\frac{S_a(t)}{S_\cdot(t)}}$$

where:

$S_a(t)$ is the total of the symptomatic variable for the small area "a" at time "t".

$S_\cdot(t)$ is the corresponding total for the large area at time "t".

$P_a(t)$ is the population of the small area "a" at time "t".

$P_.(t)$ is the corresponding population for the small area "a" at time "t".

When all the quantities in the equation are known except $P_a(t+1)$, it can be calculated as:

$$P_a(t+1) = P_.(t+1) \cdot \frac{P_a(t)}{P_.(t)} \cdot \frac{S_a(t+1)}{S_.(t+1)} \cdot \frac{S_.(t)}{S_a(t)}$$

This method is also simple but risky in the assumption that there will not be fluctuations in the change in the proportion of population and the symptomatic indicator from year to year.

(e) Housing Unit

This technique establishes a relationship between the number of dwelling units and population via family-size multipliers. Dwelling units can be estimated by hydro connections, telephone connections, vacancy rates, building permits, land use surveys and other local records. Net change in the dwelling units are presumed to indicate net change in population.

The method is defined below:

$$P_a(t) = H_a(t) \times PPH_a(c) + GRP_a(t)$$

where

$P_a(t)$ = Population for small area "a" at time "t"

$H_a(t)$ = The number of households for small area "a" at time "t"

$PPH_a(c)$ = Average number of persons per household in most recent census for small area "a"

$GRP_a(t)$ = The number of persons living in group quarters for small area "a" at time "t".

This method is not complex. One of its drawbacks is that it is a sequential method; the number of occupied households is estimated, then the average size of households. Each step in the estimation process requires decision making that commonly must be based on sketchy information or intuition. This method does not give compositional detail for the population being estimated.

4.3 REGRESSION SYMPTOMATIC TECHNIQUE

4.3.1 Introduction

The regression symptomatic techniques (ratio and difference correlation) are used to produce small area estimates using a set of symptomatic indicators as predictors. In recent years, this method has become very popular to estimate population for small areas in Canada, U.S.A., Australia and New Zealand.

4.3.2 Description

The method involves the development of a regression model using Census data, and auxiliary or symptomatic variables which are available at the small area level on an ongoing basis.

A regression equation of the following form is built:

$$\Delta P = a + b_1 \Delta x_1 + b_2 \Delta x_2 + \ldots + b_n \Delta x_n + \varepsilon$$

ΔP represents the change in the small area's proportion of the total for the variable of interest (eg population, unemployment).

Δx_i represents the change in the proportion of a symptomatic variable for the small area.

The difference correlation method uses differences to represent the changes in the variables while the ratio correlation method uses ratios to represent the changes.

Using data from two recent censuses the parameters a, b_1, b_2, ..., b_n can be estimated. Then estimated parameters can be combined with knowledge of the changes in the symptomatic variables to produce post-censal estimates.

Under the ratio correlation model, the estimate is:

$$\hat{Y}_a(t+1) = (P_a(t) * \Delta P_a) * \hat{Y}_\cdot(t+1)$$

Under difference correlation:

$$\hat{Y}_a(t+1) = (P_a(t) + \Delta P_a) * \hat{Y}_\cdot(t+1)$$

where $\hat{Y}_a(t+1)$ is the estimate of interest for small area "a".

$P_a(t)$ is the known proportion of the total for the variable of interest contributed by small area "a", at time "t".

$\hat{Y}_\cdot(t+1)$ is the known current total for the variable of interest.

$\Delta P_a = \hat{a} + \hat{b}_1 \Delta x_{a_1} + \ldots + \hat{b}_n \Delta x_{an}$ where $\hat{a}$, $\hat{b}_1$, ... $\hat{b}_n$ are the estimated parameters of the linear model and Δx_{a_1}, ..., Δx_{an} are the changes in the symptomatic variables for area "a" between time "t" and "t+1".

4.3.3 Discussion

Schmitt and Crosetti and many others have claimed that the ratio-correlation method is one of the most accurate methods for preparing population estimates (Schmitt and Crosetti, 1959; Goldberg and Balkrishnan, 1960; Goldberg, Rao and Namboodiri, 1964; Swanson, 1978; N.R.C., 1980; Mandell and Tayman, 1982). The accuracy of population estimates is generally measured by

the mean absolute percent error of the estimate (MAPE). Later, some researchers including Schmitt and Grier suggested that the difference-correlation method is an improvement over the ratio-correlation method (Schmitt and Grier, 1966; O'Hare, 1976). This is because the difference-correlation method produces a lower mean square error (M.S.E.), higher correlation between the dependent and independent variable with a resulting higher coefficient of multiple correlation (R^2). These features are often used to evaluate the fitting of a regression model and are considered desirable.

However, no consistent relationship between the higher R^2 and the mean absolute per cent error has been observed. The accurary of population estimate produced by the regression method is highly dependent on the temporal stability of the regression coefficients. In this respect, a recent study has shown that the ratio-correlation method was more suitable then the difference-correlation method (Mandell and Tayman, 1982). The difference-correlation method shows higher instability in the regression coefficients over time periods (Spar and Martin, 1979).

A review of both techniques has revealed that neither the ratio-correlation, nor the difference-correlation method uniformly or routinely outperforms the other (O'Hare, 1980; Verma, Basavarajappa and Bender, 1982). Thus, a choice of ratio or difference-correlation is dependent on a thorough evaluation of the performance of the regression method based on past data.

As mentioned earlier, the method is highly effective at the small area level. It provides detail on spatial distribution of population. This becomes weak, however, in the case of an area currently experiencing a growth rate either much faster or much lower than was observed in the past.